Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 16:54:58
REFERENCE NO: 297

Author Names & Affiliations

- Claire Paris - Rosenstiel School of Marine and Atmospheric Science, University of Miami
- Ben Kirtman - RSMAS, University of Miami
- Natalie Perlin - Center for Computational Science (CCS), University of Miami
- Ana Vaz - RSMAS, University of Miami
- Igal Berenshtein - RSMAS, University of Miami

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Ocean Science, Oceanography, Dispersion and Migration

Title of Submission

Connectivity Modeling System: A Virtual Tracking Framework for Interdisciplinary Lagrangian Applications

Abstract (maximum ~200 words).

A common interdisciplinary challenge is the accurate prediction of the 3D motion trajectories of properties, organisms, and particles in fluids at small (millimeters) and large (kilometers) scales. Particles are transported through both the aquatic and the atmospheric media, varying from biotic organisms (e.g. microbes, larvae) to abiotic particles (e.g. oil, debris). Their transport dynamics are governed by the interplay between the transported objects' physical-chemical-biological characteristics and those of their environment, which numerical modeling requires many task computing.

In the past decade we have developed a integrated approach to this complex problem by creating an open-source toolbox, the Connectivity Modeling System (CMS). The software system has been applied worldwide in numerous globally important aspects. The code has a multidisciplinary relevance, and could become a national resource for Lagrangian applications. It is well-documented and has a transparent structure facilitating customization, for both public-domain use and custom-built applications.

However, there are a number of factors that currently hinder the CMS from fulfilling its potential of becoming nationally operational tool used by communities. These limitations could considerably be resolved with a comprehensive technical and scientific approach to advance the structural development of the resources and cyberinfrastructure.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

A common challenge among a range of disciplines is the accurate prediction of the three-dimensional motion of properties, organisms and/or particles in fluids at various scales (millimeters to kilometers) and media (air and water).

To adequately estimate dispersal and transport of “particles” (e.g., particulate matter, debris, larvae, contaminated water parcels, oil and other pollutants), the motion of the surrounding environment (advection), the inherent motion of the particles themselves, and their state/fate must be computed (e.g. of chemical compounds, larvae). The complexity of this problem requires the collaboration between different disciplines, as well as the technological development , which enable to address the peculiarities of each scenario.

Physical tagging of small particles or organisms, and a direct measurement of their movement is impractical and cost prohibitive at large patio-temporal scales. The increasingly popular method to address these issues is to use computer models to solve the motions of fluids, and particles embed on them, where transport is addressed “offline” using a Lagrangian stochastic model (LSM) applications. In many LSMs, hundred millions of abiotic or biotic particles are “released” to the ocean, oftern advected with the currents for long times (years), along a very long distance (thousands of kilometers; Fiksen et al. 2007). Tracking particles for longer time periods, from shallow coastal areas to the deep ocean and back, is typically solved with a tradeoff between domain extent and spatial resolution. It thus becomes critical to use simultaneously several distinct circulation models to evaluate coastal and ocean scale conditions in order to improve the accuracy of Lagrangian predictions. These requirements are often computationally complex and very demanding, with large data management efforts that can limit the scope of the applications.

Over the last decades, technological development has allowed computer models to become more complex and accurate. This has opened up possibilities for the investigation of intricate systems and interactions. A substantial amount of advance has been made to understand the mechanisms underlying transport in the atmosphere, oceans and other aquatic systems, and their effects on climate, ecosystems and economy. However, to move forward, we need to increase the spatial-temporal scales at which we look at transport, while also increasing the precision to solve physical-chemical and biological traits of individual particles. The solution for the present constrains will entail the participation of experts from distinct disciplines, as well as the development of a robust and efficient system software and applications. With the growing investment in interdisciplinary research approaches, Such collaboration can be highly rewarding since, on one hand: the refinement of the biological/ chemical/ physical aspects implemented in CMS can improve its prediction accuracy, the implementation of these properties in the model allow researchers from these fields to examine the broad-scale effects of small-scale processes which they study.

Considering this complex problem, we have developed over the past 2 decades an open-source Fortran toolbox, the Connectivity Modeling System (CMS, Paris et al., 2013). The Lagrangian framework and inter-disciplinary code is unique with a series of standalone modules for the tracking of biotic and abiotic particles in the ocean. CMS includes ocean dynamic, inertial motion, pollutants, plankton, etc. In addition to the traditional Lagrangian trajectory output of LSM, CMS direct output consist of 3D connectivity matrices, defining the probability of connections between source and sink location of particles (e.g. micro-plastics, debris, larvae, etc.) not only in the horizontal plane but also in the vertical (Vaz et al. 2016) and their connectivity in a geographical network perspective at multiple spatial and temporal scales (Holstein et al. 2014). The tool is inherently multi-scale, allowing for the seamless moving of particles between grids at different resolutions. The CMS has been used on velocity fields from OFES, HYCOM, NEMO, MITgcm, UVic, ECCO2, SOSE, MOM and many other coastal and ocean general circulation models (OGCMs) to compute dispersion, connectivity, and fate in applications including large scale oceanography, marine reserve planning, movement of marine biota, and drift of plastic all across the world. When compared to other offline Lagrangian applications, CMS is very accurate since it uses RK4 time stepping and an adaptive tri-cubic interpolation of the OGCM variables both in time and space. It is designed to be modular, meaning that it is relatively easy to add additional physical forcing, interactions, and behaviors on the particles. Modules distributed with the code include mixed layer dynamics, random walk and correlated random walk diffusion, mortality, diel vertical migration, buoyancy, inertial movement, orientation, mass spawning, 3D settlement habitat, backtracking etc. Moreover, CMS has a probabilistic approach not only implementing time-variant ensemble model runs, but whereby particle attributes are taken at random from a distribution of traits. These unique features of CMS are essential to increase the accuracy of predictability as well as to the assessment of uncertainties (DeGwou et al. 2011, Paris et al. 2012, Le Henaff et al. 2012, Socolofski et al. 2015). CMS has been used worldwide for numerous applications, including for the Deepwater Horizon oil spill, the Fukushima Daiichi nuclear disaster, and plastic gyres in the ocean.

We have thus implemented a multi-component application, which consists of many individual tasks that utilize a distributed set of computational and data management resources. However, the requirements for these applications are often computationally complex and demand large data management efforts. These HPC challenges create serious scientific limitations (please see limitations in Q2).

NOTE: The CMS has an estimate of >1000 users who downloaded the code since it was published in 2013. According to the Web of Science, at least 100 peer reviewed articles cited the CMS -more than 2 publications per month since the release of the open-source code (Paris et al. 2013). These studies range from oceanographic (Quin et al. 2014, 2015; Rossi et al. 2013), biological and spatial conservation (Nanninga et al. 2015, Snyder et al. 2015), fisheries (Karnauskas et al. 2013, Grüss et al. 2014), oil spill and plastic pollution (Aman et al. 2015; van Sebille, 2014), uncertainty analyses (Quin et al. 2014, 2015), to understanding the effect of desalinization (Zhan et al. 2015). While applications using CMS operate at all spatial and temporal scales, from global and evolutionary scales (Hellweger et al. 2014, Wood et al. 2014, Froyland et al. 2014), to individual trajectories at the meter scale and group behavior in larvae at the micro-scale (Fujimura et al.

2014), CMS can also operate across multiples scales (Holstein et al. 2014) since it has the unique capacity of Lagrangian nesting offline by computing trajectories seamlessly across models with different model grid resolutions.

Aman, Z.M. et al. (2015) High-pressure visual experimental studies of oil-in-water dispersion droplet size. Chemical Engineering Science, 127, 392-400.
DeGouw JA et al. (2011) Organic Aerosol Formation Downwind From the Deepwater Horizon Oil Spill, Science 331:1295.
Holstein DM et al. (2014) Consistency and inconsistency in multispecies population network dynamics of coral reef ecosystems. Marine Ecology Progress Series, 499, 1-18.
Froyland, G. et al. (2014) How well-connected is the surface of the global ocean? Chaos: An Interdisciplinary Journal of Nonlinear Science, 24, 033126.
Fujimura, et al. (2014) Numerical simulations of larval transport into a rip-channeled surf zone. Limnology and Oceanography, 59, 1434-1447.
Paris, C.B. et al.. (2013) Connectivity Modeling System: A probabilistic modeling tool for the multi-scale tracking of biotic and abiotic variability in the ocean. Environmental Modelling & Software, 42, 47-54.
Socolofsky SA, et al. (2015) Case Study: Intercomparison of Oil Spill Prediction Models for Accidental Blowout Scenarios with and without Subsea Chemical Dispersant Injection, Marine Pollution Bulletin 96:1-2 DOI: 10.1016/j.marpolbul.2015.05.039
Zhan,P. et al. (2015) Far-Field Ocean Conditions and Concentrate Discharges Modeling Along the Saudi Coast of the Red Sea. Intakes and Outfalls for Seawater Reverse-Osmosis Desalination Facilities (ed. by T.M. Missimer, B. Jones and R.G. Maliva), pp. 501-520. Springer International Publish

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The CMS model code has multidisciplinary relevance and application, and could become a national resource for Lagrangian modeling approach. It is well-documented and has transparent structure facilitating customization, for both public-domain use and custom-built applications. There is a number of factors that currently limit wider use of the CMS by the community, and that could be considerably improved with further model improvement and structural development of the resources and cyberinfrastructure, as outlined below. These limiting factors could be sorted into different groups as follows: 1) code optimization, 2) post-processing package development, 3) code and model results sharing challenge, maintaining high security when required, 4) providing high redundancy of the code and the results.
1. Code Optimization
In the first group of limiting factors, improvement of the scheme for the optimal computational resources reservation for different implementation methods in multi-processor or multi-thread applications (MPI, OpenMP, etc.) and a user-friendly guide to their proper use in variety of systems are primordial. Additionally, exploring and expanding the options for model output storage is necessary to overcome current issues of placing unnecessary load on the storage network in large systems due to heavy I/O operations, such as appeared to be the issue with the popular netcdf format, due to its inefficient internal algorithm of data access and file update. This limitation becomes critical when the model is used in a probabilistic mode, with numerous ensembles of initial conditions (i.e. millions of release locations and times, millions of particles), as well as long integration times and high output frequency.
2. Post-processing Package Development
The challenge is to use the output format that is self-describing with enough metadata, relatively compact, and is light on the storage network when frequent model output is required. Furthermore, developing a package of post-processing routines, as well as additional off-line conversion of model output into a variety of formats will offer an attractive solution to analyze model results, their visualization, and their preferred storage data type for the end-user.
3. Code and Model Results Sharing
The third group of factors limiting code and data sharing is essential for further code development and improvement by a community of users from various research areas, from one side, as well as the crucial need for model results exchange between the collaborators. Additionally, high security and controlled access to the data could be required by any DOD or DOE-sponsored research. Indeed, private applications that may be related to national resources, energy, or security should have their specific algorithm (or module) and model output secured. Typical block storage solutions are pending improvement and enhanced access capabilities of the object storage via the Representational State Transfer (REST interface using HTTP protocol). Current options offered by the Amazon Web Services (AWS) of S3 object storage for business customers could be a good alternative for such data access, but may impose undefined costs and not meet the

required security standards for classified data.

4. Providing High Redundancy
The fourth group of limiting factors addresses the need for the high redundancy of the stored model results and version control of the code. Production of model results and their post-processing usually involves high costs and time investment of both human- and cpu-hours, and thus high precautions have to be taken to prevent data loss due to accidental removal or equipment failure. Currently used backups or archiving of the block storage are very time-consuming and ineffective, and could be replaced by a high redundancy options offered by object storage solutions, customized automated policies for archiving and data lifecycle, similar to S3 and Glacier AWS implementations. Further development and testing the model code needs to have robust version control alternative.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Scientific background of the personnel is primary requirement to conduct quality research. In addition to that, digital proficiency and strong programming skills, as well as openness towards new technological solutions is a highly sough-after component in developing new workforce for scientific research and engineering. We, the authors of this letter, are from various scientific backgrounds, working in four departments of the university of Miami: Ocean Sciences, Atmospheric Science, Center for computation Science, and Marine Biology and Ecology. Yet, we are involved in interdisciplinary studies, working together in developing CMS. This proposed program will enhance the interdisciplinary nature of sciences and will also increase the representation of women in STEM.
We are planning to work in collaboration with NCAR to address some of these challenges. We believe that improving the cyberinfrastructure of highly interdisciplinary Lagrangian applications and its visualization capacities will advance the frontiers of science and engineering over the next decades. Ou vision is to make the CMS software system a national operational resource.

Consent Statement